Transcription Regulation and Genome Organization

Hermsen, R.

2008

Samenvatting

Een van de bijzondere eigenschappen van levende wezens is dat ze kunnen reageren op prikkels. Veel organismen nemen bijvoorbeeld veranderingen waar in lichtintensiteit (*zien*), mechanische krachten (*voelen*), en de concentratie van bepaalde moleculen in de omgeving (*ruiken* en *proeven*). Ze reageren natuurlijk niet voor niets op zulke waarnemingen: deze reacties zijn gedurende miljoenen jaren van evolutie ontwikkeld zodat organismen zich kunnen aanpassen aan veranderingen in hun omgeving, en uiteindelijk een grotere kans hebben te overleven en zich voort te planten.

Het zal geen verrassing zijn dat *mensen* kunnen zien, voelen, ruiken en proeven, en ook niet dat zij hun gedrag aanpassen naar aanleiding van hun zintuiglijke waarnemingen. Maar het is een stuk minder vanzelfsprekend hoe ééncellige organismen die niet groter zijn dan een duizendste van een millimeter, zoals bacteriën, kennis over allerlei fysisch-chemische grootheden kunnen verkrijgen, afwegen en uitbuiten. Toch doen ze dat. Bacteriën zwemmen op voedsel af, zoeken een warm plekje op, zetten hun biologische klok gelijk, beoordelen de dichtheid van hun kolonie, meten de osmotische druk van hun omgeving en houden in de gaten welke soorten suikers er voorhanden zijn — om maar een paar dingen te noemen.

Transcriptieregulatie is één van de belangrijkste mechanismen die cellen gebruiken om te reageren op hun omgeving, en bovendien het onderwerp van dit proefschrift. De Hoofdstukken 2 en 3 gaan over transcriptieregulatie in bacteriën. We onderzoeken de mechanismen van transcriptieregulatie en wat hun mogelijkheden en beperkingen zijn. In de Hoofdstukken 4, 5 en 6 beschouwen we hoe transcriptieregulatie de indeling van genomen beïnvloedt. Alle analyses zijn gebaseerd op theoretische modellen, maar we gebruiken experimentele gegevens om deze modellen en hun voorspellingen te testen. We concentreren ons voornamelijk op prokaryoten (organismen die geen celkern hebben), met de darmbacterie *Escherichia coli* in de hoofdrol — maar voor verschillende andere organismen is een bijrol weggelegd.

Transcriptie en transcriptieregulatie

Cellen nemen veel van hun beslissingen op het niveau van *transcriptie*. Transcriptie is het moleculaire proces waarbij genen (stukken DNA die de instructies bevatten voor het maken van één enkel type eiwit[7]) worden gekopieerd (getranscribeerd). Dit kopieerproces wordt uitgevoerd door een moleculaire machine die RNA-

[7] Sommige genen coderen voor een stabiel RNA-molecuul in plaats van een eiwit.

polymerase heet (RNAp). Het kopie heeft een net andere structuur dan DNA en wordt mRNA genoemd. Elk mRNA-molecuul wordt vervolgens gebruikt als blauwdruk voor het maken van een specifiek eiwit. Daardoor bepaalt de frequentie waarmee de transcriptie van een bepaald gen plaatsvindt grotendeels hoeveel kopieën van het corresponderende eiwit in de cel aanwezig zijn.

Transcriptieregulatie is het proces waarmee cellen reguleren hoe vaak bepaalde genen worden getranscribeerd. Indirect worden hiermee dus de eiwitconcentraties in de cel gereguleerd. Dit is heel belangrijk, omdat eiwitten allerlei eigenschappen van de cel bepalen. Veel eiwitten functioneren bijvoorbeeld als enzymen, die de snelheid van chemische reacties in de cel beïnvloeden. Andere eiwitten zijn de bouwstoffen waaruit de meeste structuren in de cel zijn opgebouwd. En weer andere eiwitten werken als minuscule motortjes die chemische energie omzetten in mechanische beweging. Kortom, door de transcriptiesnelheid van groepen van genen te reguleren, kunnen cellen hun samenstelling drastisch aanpassen en daardoor ook hun gedrag en vorm.

Bij de regulatie van transcriptie speelt een speciale familie van eiwitten een grote rol. Deze eiwitten heten *transcriptiefactoren*. Ze functioneren doordat ze kunnen binden aan specifieke stukken DNA. Deze "parkeerplaatsen" zijn meestal vlak bij de beginpunten van genen te vinden, daar waar de RNAp begint met het transcriptieproces. Wanneer transcriptiefactoren binden aan hun bindingsplaatsen kunnen ze de efficiëntie van de eerste stappen van het transcriptieproces beïnvloeden en daardoor ook de frequentie veranderen waarmee transcriptie plaatsvindt.

Logische beslissingen

Bacteriën moeten vaak *logische* beslissingen nemen. Een beroemd voorbeeld is de manier waarop *E. coli* zijn maaltijd kiest. *E. coli* haalt zijn energie uit suikers; maar om de verschillende soorten suikers, zoals glucose, lactose, galactose en arabinose te kunnen opnemen en verteren, moet hij een aantal eiwitten (enzymen) produceren die de benodigde stofwisselingsreacties stimuleren. De verschillende suikers zijn meestal niet allemaal aanwezig in de omgeving van de bacterie. Omdat de productie van de eiwitten onder andere energie kost, zou het niet erg efficiënt zijn om ze continu aan te maken. Daarom beslist *E. coli* afhankelijk van de beschikbaarheid van de suikers welke enzymen hij wel of niet wil produceren.

In feite gaat *E. coli* nog een stapje verder. Omdat hij het snelst kan groeien als hij glucose eet, is dat zijn favoriete maaltijd. Daarom produceert *E. coli* de eiwitten die nodig zijn voor de vertering van, zeg, lactose, enkel als er lactose beschikbaar is en geen glucose. Dit laat zien dat de bacterie meerdere gegevens — de beschikbaarheid van de verschillende suikers — betrekt bij zijn beslissing. De besluitprocedure is, kort gezegd: maak eiwitten LacY, LacZ en LacA alleen als er lactose is EN NIET glucose. De relatie EN NIET is een voorbeeld van een Boolese functie (zie Kader 1 voor meer uitleg).

Deze besluitprocedure wordt met behulp van transcriptieregulatie uitgevoerd. Dat werkt als volgt. *E. coli* bevat een speciale transcriptiefactor, genaamd LacI.

Kader 1: Boolese logica en logische poorten

In veel wetenschapsgebieden wordt een speciale indeling van logische redeneringen gebruikt. Deze wordt de Boolese logica genoemd, naar de 19e-eeuwse Britse wiskundige en filosoof George Boole, die er de grondlegger van is.

We kunnen de Boolese logica introduceren aan de hand van een paar voorbeelden. Sommige beslissingen zijn van het type A EN B. Hier staan A en B voor willekeurige uitspraken. Een goed voorbeeld is als iemand zegt: "Ik ga alleen zeilen als het zonnig weer is en het bovendien waait". Als we de uitspraak "Het is zonnig weer" nu A noemen en de uitspraak "Het waait" aanduiden met B, dan heeft de beslissing om te gaan zeilen inderdaad de vorm "Ik ga zeilen als A EN B". Een andere uitspraak kan zijn "Ik ga alleen fietsen als het zonnig weer is en het bovendien *niet* waait". Deze zin is van het type A EN NIET B. Op deze manier bestaat er een hele verzameling aan Boolese functies, waaronder EN, OF, EN NIET, OF NIET, etcetera.

Een apparaatje dat een van de Boolese beslissing kan uitvoeren, heet een *logische poort*. Zulke logische poorten vormen de basis van alle digitale elektronica. Het voorbeeld in de tekst over de maaltijdkeuze van *E. coli* laat zien dat bacteriën blijkbaar een mechanisme bevatten dat kan functioneren als een logische poort. Eigenlijk is de bacterie dus een soort analoog computertje.

Deze transcriptiefactor bindt in de afwezigheid van lactose op een speciale plek aan het DNA. Daardoor verhindert hij de transcriptie van de genen die coderen voor de eiwitten LacY, LacZ en LacA; deze eiwitten worden in afwezigheid van lactose dus niet geproduceerd. Maar, als er wel lactose in de omgeving is, dan bindt lactose aan de transcriptiefactor LacI[8]. LacI verandert hierdoor van vorm; in deze vorm bindt het veel slechter aan het DNA en blokkeert het de transcriptie niet langer. De transcriptiefactor CRP doet iets soortgelijks, maar dan voor glucose: CRP activeert transcriptie als er geen glucose aanwezig is. Op deze manier worden de lactose-genen alleen gebruikt als er lactose is maar geen glucose.

Mechanismen van transcriptieregulatie

Door middel van ingewikkelde experimenten zijn een aantal mechanismen ontdekt die bacteriën gebruiken om de transcriptiesnelheid te reguleren. De meeste van deze mechanismen werken doordat transcriptiefactoren andere moleculen helpen bij hun binding aan het DNA, of precies andersom, ze daarbij hinderen. Bijvoorbeeld, als een transcriptiefactor de RNAp helpt bij het binden aan zijn bindingsplaats (deze wordt de *promoter* genoemd), dan wordt de transcriptie geactiveerd. Als, aan de andere kant, de transcriptiefactor de binding van RNAp blokkeert, dan wordt transcriptie onderdrukt.

In de Hoofdstukken 2 en 3 van dit proefschrift bestuderen we welke types

[8]Dit is niet helemaal waar: eigenlijk bindt niet lactose aan LacI, maar allolactose. *E. coli* zet lactose om in allolactose.

Kader 2: Evolutionaire algoritmes

Biologische evolutie is het samenspel van mutaties en (natuurlijke en seksuele) selectie. Mutaties zorgen er voor dat verschillende individuen in een populatie niet precies hetzelfde genoom hebben. Individuen die door hun genoom beter in staat zijn zich voort te planten dan anderen, krijgen gemiddeld meer nakomelingen; genen die de voortplantingskansen verhogen hebben dus grote kans om in de loop van de tijd in steeds grotere aantallen voor te komen. Door miljoenen rondes van mutaties en selectie kunnen zo organismen ontstaan die heel goed zijn aangepast op hun omgeving. In de natuur heeft het proces van evolutie in de loop der miljoenen jaren heel geavanceerde organismen opgeleverd.

Een evolutionair algoritme is een type computerprogramma dat een evolutionair proces nabootst om oplossingen te vinden voor een bepaald complex (ontwerp)probleem. De computer slaat eerst een groot aantal willekeurige ontwerpen in zijn geheugen op. Die eerste, willekeurige ontwerpen zijn in het algemeen heel slechte oplossingen voor het betreffende probleem. Vervolgens kiest het programma de beste ontwerpen uit, kopieert ze een aantal keer, en brengt er hier en daar willekeurige wijzigingen in aan. Sommige ontwerpen zijn er waarschijnlijk slechter op geworden, maar een paar zijn wellicht iets verbeterd. Weer selecteert het programma de beste ontwerpen. Na heel veel rondes van muteren en selecteren kan zo een heel geavanceerde ontwerp worden geconstrueerd. Een leuk aspect aan deze methode is dat je op deze manier iets ingewikkelds kunt ontwerpen, zonder dat je de oplossing zelf hoeft te bedenken.

Wij gebruikten een evolutionair algoritme om DNA-sequenties te vinden die als logische poorten kunnen functioneren. Het algoritme vond vaak oplossingen die we zelf nog niet hadden bedacht.

beslissingen in principe kunnen worden geïmplementeerd met de mechanismen die bekend zijn. Om dat uit te zoeken, formuleren we een kwantitatief model van transcriptieregulatie. We combineren dit model met een evolutionair algoritme basale om transcriptieregulatiesystemen te ontwerpen die een door ons gekozen functie vervullen. Zo kunnen we het scala aan mogelijke ontwerpen verkennen. In Kader 2 is meer te lezen over evolutionaire algoritmes.

Het blijkt dat de eenvoudige mechanismen van transcriptieregulatie enorm veelzijdig zijn. Met behulp van vrij complexe patronen van bindingsplaatsen kunnen alle mogelijke logische beslissingen met twee input-signalen worden geïmplementeerd. De beste ontwerpen bestaan uit modules van bindingsplaatsen die allemaal direct naast elkaar liggen. De transcriptiefactoren die aan deze plaatsen binden, helpen elkaar bij het binden. Dit coöperatieve gedrag leidt tot een scherpe reactie van de transcriptiefrequentie als functie van de concentraties van de transcriptiefactoren. Meer geavanceerde effecten kunnen worden bereikt als de modules (gedeeltelijk) met elkaar overlappen. Dat introduceert competitie op het niveau van modules, die immers niet tegelijk gebonden kunnen zijn. Welke module domineert, kan in zulke situaties sterk afhangen van de concentraties van

de verschillende transcriptiefactoren. Dit kan worden uitgebuit om verschillende signalen tegen elkaar af te wegen.

De mogelijkheden worden nog meer vergroot als we terugkoppeling (feedback) toelaten. In het meest eenvoudige geval codeert het gereguleerde gen voor een transcriptiefactor die op zijn beurt zijn eigen transcriptiefrequentie reguleert. Dit heet auto-regulatie. Het is aangetoond dat dit het mogelijk maakt om de dynamische eigenschappen van de systemen af te stellen — bijvoorbeeld, de gevoeligheid voor ruis of de reactie-snelheid. Onze resultaten tonen aan dat auto-regulatie ook een efficiënter repressiemechanisme mogelijk maakt en alternatieve mechanismen biedt voor het integreren van signalen. De mechanismen die we vinden werpen een nieuw licht op de mogelijke functies van feedback-systemen in transcriptieregulatie.

Chromosoom-organisatie

De processen van transcriptie en transcriptieregulatie hebben een grote invloed op de manier waarop genen verdeeld zijn over chromosomen. Alle stukken DNA die een rol spelen in transcriptieregulatie, zoals de bindingsplekken voor RNAp en transcriptiefactoren, nemen plaats in beslag op het DNA. Daardoor beïnvloeden ze de afstanden tussen genen. Inderdaad zien we dat deze sequenties direct vóór en na genen hun sporen achterlaten op de kansverdelingen van afstanden tussen genen. Omgekeerd onthullen de statistische eigenschappen van de afstanden tussen genen allerlei informatie over de regulatiemechanismen die door het organisme worden gebruikt.

Afstanden tussen genen

Om de kansverdelingen van de afstanden tussen genen goed te bestuderen, vergelijken we ze met modellen. We maken er gebruik van dat, wiskundig gezien, deze modellen precies overeenkomen met modellen van één-dimensionale gassen. In deze analogie komen genen overeen met gasdeeltjes en het DNA speelt de rol van een één-dimensionale, eindige ruimte.

Het beste model is het Constantekrachtmodel. In dat model nemen we aan dat de genen worden vergezeld door sequenties ten behoeve van de regulatie die plaats innemen en daarom de genen als het ware uit elkaar houden. De genen zijn willekeurig verdeeld, behalve dat ze zelden overlappen en elkaar op korte afstanden “afstoten”. Dit model komt erg goed overeen met de verdelingen in organismen zoals *E. coli* en *Saccharomyces cerevisiae* (bakkersgist).

De genoomdata wijken op verschillende punten af van het Constantekracht-model. Deze afwijkingen leiden tot interessante biologische voorspellingen. Bijvoorbeeld, in de meeste *schimmels* heeft de kansverdeling van afstanden tussen divergente genen — naburige genen die in tegengestelde richting en in divergente oriëntatie worden afgelezen — twee pieken, wat sterk suggereert dat deze genomen veel bi-directionele promoters bevatten. Net zoiets is het geval in *E. coli*: we vinden een flink overschot aan convergente gen-paren — naburige genen die in te-

Kader 3: Promoters en terminators, hoofdletters en punten

Je kunt je DNA voorstellen als een lange reeks letters, net als een tekst. In die vergelijking is een gen zoiets als een *zin*.qpenvaosvkHet verwarrende is dat tussen de genen op het DNA ook veel letters staan die geen betekenis hebben. ·worden gescheiden Dit is allemaal niet zo'n probleem, omdat er codes op het DNA staan die vertellen waar een gen begint en waar hij eindigt. Het beginsignaal heet een *promoter* en kan het best vergeleken worden met een hoofdletter, die immers het begin van een zin aangeeft. Het eindsignaal is een *terminator* en heeft dezelfde functie als een punt.

Bij het bekijken van de afstanden tussen genen van *schimmels* vonden we aanwijzingen dat veel promoters bi-directioneel kunnen zijn. Dat wil zeggen dat ze twee kanten op functioneren. ·illustreren. Zoals u ziet, gebruikt deze zin dezelfde hoofdletter als de vorige zin. Dat kan natuurlijk alleen maar omdat beide zinnen met een heel speciale letter beginnen die ook op z'n kop bruikbaar is.

In *Escherichia coli* vinden we juist sterke aanwijzingen voor veel bi-directionele terminators.·uiz Deze zin gebruikt dezelfde punt als de vorige zin Dat kan omdat een punt er omgekeerd precies hetzelfde uitziet als rechtop. Dat geldt ook voor bi-directionele terminators; deze sequenties zijn bij benadering palindromen (sequenties die symmetrisch zijn, waardoor je ze in twee richtingen kunt lezen, zoals het woord "meetsysteem").

gengestelde richting en in convergente oriëntatie worden afgelezen — die bijzonder dicht bij elkaar verblijven; we voorspellen dat deze gen-paren een bi-directionele terminator delen (zie Kader 3).

Operons

Een speciale eigenschap van de meeste (zo niet alle) prokaryoten en een paar eukaryoten is dat hun genen zijn georganiseerd in zogenaamde *operons*. Een operon is een kluster van genen die samen worden getranscribeerd tot één mRNA. Genen in een operon bevinden zich meestal heel dicht bij elkaar en worden ook in dezelfde richting afgelezen; ze hebben een zogenaamde *tandem* oriëntatie. Daardoor bestaat de verzameling van tandem gen-paren in zulke genomen uit twee groepen: de gen-paren die in hetzelfde operon zitten, en de paren die in verschillende operons zitten. De sequenties die tussen deze genen in liggen bevinden zich daardoor ofwel *in* een operon of *tussen* twee operons. Deze tweedeling is ook zichtbaar in de verdeling van afstanden tussen tandem genen: deze is grotendeels consistent met ons model, behalve dat een duidelijk overschot aanwezig is op korte afstanden. Door dat overschot verraadt de verdeling van afstanden de aanwezigheid van operons.

De vraag *waarom* genen zijn georganiseerd in operons, is een onderwerp van continu debat. De meningen zijn grofweg verdeeld in twee kampen. Het eerste kamp betoogt dat operons gebruikt worden om genen te co-reguleren. Als een

aantal genen in een gecorreleerde manier tot expressie moeten worden gebracht — bijvoorbeeld omdat ze een gerelateerde functie hebben — dan kan dit inderdaad worden bewerkstelligd door ze in één operon te plaatsen. Het andere kamp is van mening dat de formatie van operons het gevolg is van de "horizontale overdracht" van genen: het feit dat genen soms worden overgebracht van het ene naar het andere organisme. Operons die verschillende of zelfs alle benodigde genen bevatten voor een bepaalde functie, zouden een grotere kans kunnen hebben om succesvol te worden overgebracht naar andere organismen dan losse genen. Daarom zouden operons "egoïstische" structuren kunnen zijn: hun bestaan zou dan het gevolg zijn van hun succesvolle verspreiding en niet zozeer van hun toegevoegde waarde voor het organisme.

De twee genoemde argumenten hebben een ding gemeenschappelijk: ze nemen beiden stilzwijgend aan dat operons niet zouden bestaan in afwezigheid van enige selectiedruk om ze te creëren. In Hoofdstuk 5 beargumenteren wij precies het omgekeerde: zelfs als operons geen enkel selectief voordeel met zich meebrengen—noch op het niveau van het organisme, noch op het niveau van het operon zelf—dan nog zijn operons te verwachten. De reden is dat twee tandem buurgenen van nature in hetzelfde operon zijn, *tenzij* er zich tussen hen in een terminatorsequentie bevindt. Dit betekent dat, in zekere zin, operons de "default"-indeling zijn: alleen als er voldoende en aanhoudende evolutionaire druk bestaat om de genen onafhankelijk te reguleren, kan men verwachten dat in de loop van de evolutie terminators en onafhankelijke promoters ontstaan. Tegelijkertijd worden bestaande terminators continu op de proef gesteld door allerlei soorten mutaties. Op evolutionaire tijdsschalen zullen zij enkel overleven als ze constant onder voldoende selectiedruk staan. Wanneer dat niet het geval is, zal de terminator verdwijnen en een operon ontstaan.

Om aan te tonen dat dit concept hout snijdt, presenteren we een eenvoudig model voor de evolutie van genomen en ontwikkelden we een nieuw simulatieschema gebaseerd op het wetenschapsgebied van de populatie-genetica. In simulaties van dit model ontstaan inderdaad spontaan operons en gedeelde terminators. Bovendien reproduceert het model de verdeling van genen in de organismen *E. coli* en *Bacillus subtilis*, inclusief de karakteristieke clustering van genen in operons en de verschillen in de afstanden tussen convergente, divergente en tandem genparen. *En passant* verklaart het ook waarom promoters en terminators zich in het algemeen heel dicht bij het bijbehorende gen bevinden.

De evolutie van afstanden tussen genen

Op evolutionaire tijdsschalen groeien en krimpen de afstanden tussen genen als het gevolg van invoegingen en verwijderingen van stukjes DNA. In de regio's tussen genen zijn deze stukjes typisch heel kort. Aangezien het vóórkomen van deze mutaties een kansproces is, zou je verwachten dat de lengtes van deze regio's op de lange duur een zogenaamde "random walk" beschrijven. In Hoofdstuk 6 beschrijven we een stochastisch model voor deze evolutionaire "diffusie" van sequenties tussen genen.

Dit idee kan getest worden met behulp van gegevens over gerelateerde organismen. Direct nadat twee soorten ontstaan uit een gezamenlijke voorouder, zouden de regio's tussen genen in beiden organismen even lang moeten zijn. Maar in de loop der tijd zullen invoegingen en verwijderingen ervoor zorgen dat de afstanden gaan verschillen. We kunnen ons model dus testen door de afstanden tussen genen in twee gerelateerde organismen te vergelijken met berekeningen aan het model. We vergelijken ons model met de gegevens van *Escherichia coli* en *Salmonella enterica subsp. enterica serovar Typhi*.

Het model kan ook worden gebruikt om te berekenen wat er gebeurt als een operon opsplitst of wanneer twee operons samensmelten. Door berekeningen aan dit proces kunnen we regio's opsporen waarin wellicht recentelijk een samenvoeging of splitsing heeft plaatsgevonden.